



Exceptional service in the national interest

UNLOCKING THE POWER OF LARGE LANGUAGE MODELS: PRACTICAL APPLICATIONS IN SCRIPTS AND PROGRAMS

Anthony Garland

OUTLINE

- What are LLMs
 - Basic LLM 101
 - Recent advances
- LLMs as part of a system
 - Parts
 - Memory
 - Tools
 - Database
 - LLM
 - Common design patterns
 - Chain of thought /Tree of thought
 - QA over docs
 - Agents

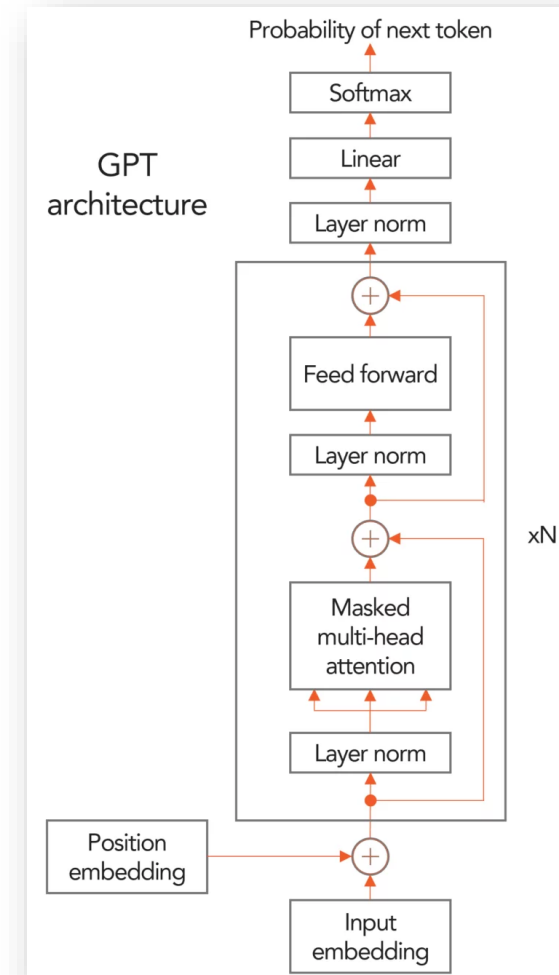
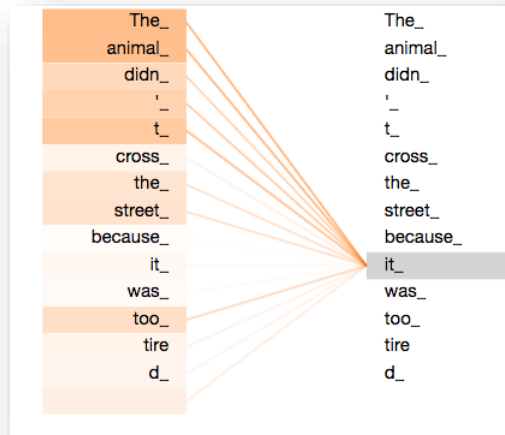




WHAT ARE LLMS?

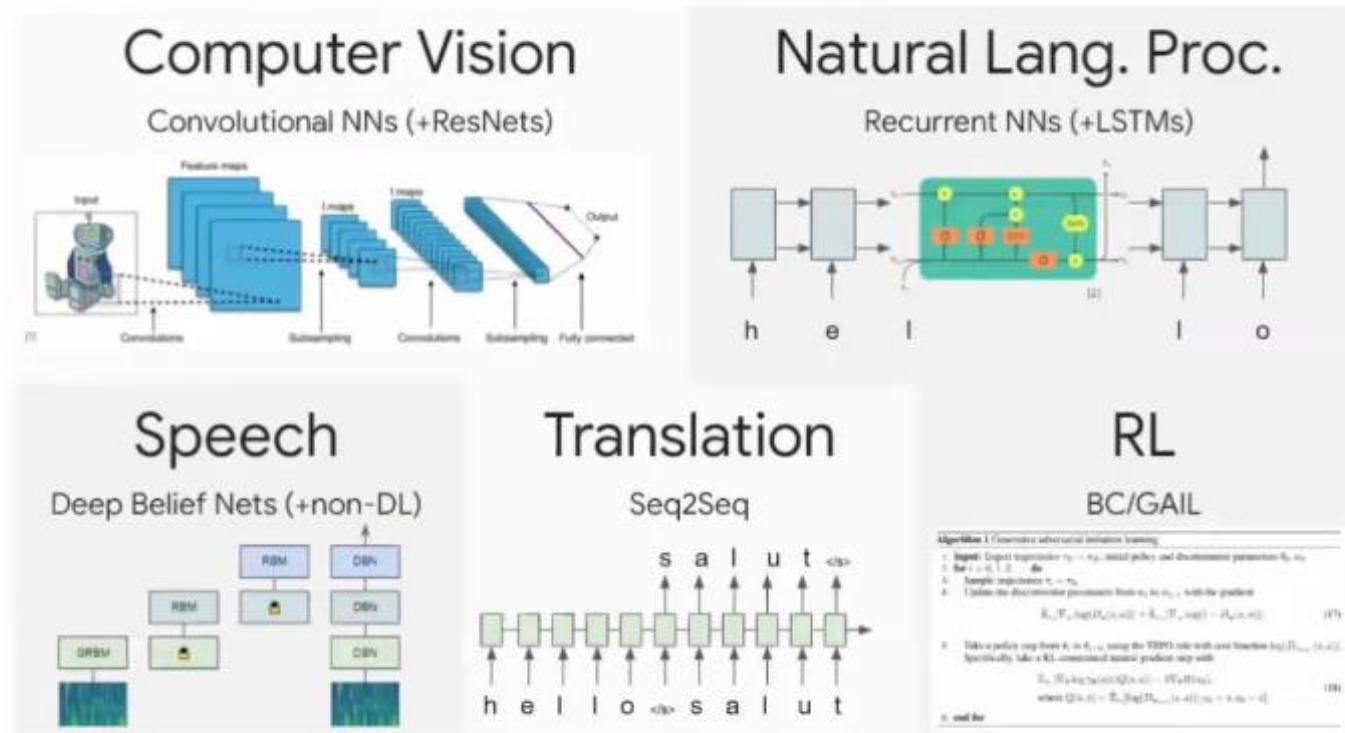
WHAT ARE LLMS

- Modern LLM are stacked layers of transformers
- Transformers
 - Convert input to a vector (embedding)
 - Use self-attention
 - Scaled dot-product
 - Each word can determine determines how much to pay “attention” to all the other input words
 - Multiple heads. Perform self-attention M times
- Stack transformer layers together N times
- Transformers are a very generic computing paradigm that can use any type of input



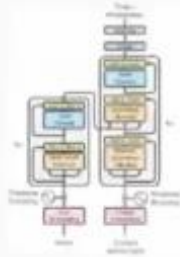
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

PRE-TRANSFORMER ML LANDSCAPE

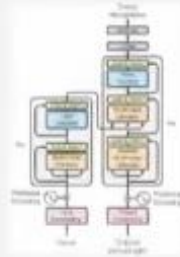


POST TRANSFORMER ML LANDSCAPE

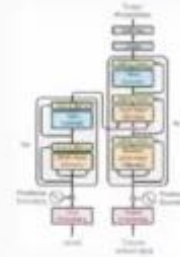
Computer Vision



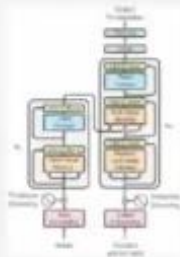
Natural Lang. Proc.



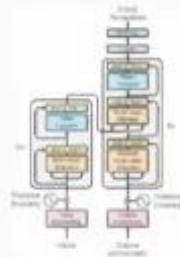
Reinf. Learning



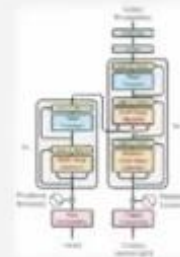
Speech



Translation



Graphs/Science



WHY INSTRUCT GPT AND CHATGPT WERE INTERESTING

- GPT3 had been around for a while, but it was hard to use.
- Useability!
 - **Alignment** of the output to be helpful to humans
 - The Chat UI worked well.

Example: Write some code to do X

How to get GPT3 to do what you want

You are GPT-3, and you can't do math.

You can do basic math, and your memorization abilities are impressive, but you can't do any complex calculations that a human could not do in their head. You also have an annoying tendency to just make up highly specific, but wrong, answers.

So we hooked you up to a Python 3 kernel, and now you can execute code. If anyone gives you a hard math problem, just use this format and we will take care of the rest:

Question: `{{Question with hard calculation.}}`

```
```python
{{Code that prints what you need to know}}
```
```

```
```output
{{Output of your code}}
```
```

Answer: `{{Answer}}`

Otherwise, use this simpler format:

Question: `{{Question without hard calculation}}`

Answer: `{{Answer}}`

Begin.

Question: What is $37593 * 67$?

```
```python
print(37593 * 67)
```
```

```
```output
2518731
```
```

Answer: 2518731

Question: `{question}`

How to get ChatGPT to do what you want



write the python code to calculate $375931 * 589$



Certainly! Here's a Python code snippet to calculate the product of 375931 and 589:

python

Copy code

```
result = 375931 * 589
print(result)
```


RECENT ADVANCES IN LLMS

- The “Large” can mean many things
 - Large # parameters ($\geq 7B$ params [Bits and Bytes paper])
 - Large training data
 - Large in quality of data
- Thriving open source community
- Highly optimized training
- Quantization for running on CPU (or less VRAM on GPU)
 - Use int4 instead of float16 datatypes
- Benchmarks are helpful, but struggling to keep up


Open LLM Leaderboard

The Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released.
 Anyone from the community can submit a model for automated evaluation on the GPU cluster, as long as it is a Transformers model with weights on the Hub. We also support evaluation of models with delta-weights for non-commercial licensed models, such as LLaMA.

Search your model and press ENTER...

LLM Benchmark (lite) Extended view About

| Model | Average | ARC (25-q) | HellaSwag (10-q) | MMLU (5-q) | TruthfulQA (MC) (8-q) |
|--|---------|------------|------------------|------------|-----------------------|
| tiiuae/falcon-40b-instruct | 63.2 | 61.6 | 84.4 | 54.1 | 52.5 |
| tindettmrs/guanaco-65b-merged | 62.2 | 60.2 | 84.6 | 52.7 | 51.3 |
| CalderaAI/398b-Lazarus | 60.7 | 57.6 | 81.7 | 45.2 | 58.3 |
| tiiuae/falcon-40b | 60.4 | 61.9 | 85.3 | 52.7 | 41.7 |
| tindettmrs/guanaco-33b-merged | 60 | 58.2 | 83.5 | 48.5 | 50 |
| austroba/llama-30b-supercoot | 59.8 | 58.5 | 82.9 | 44.3 | 53.6 |
| huddy/llama-llama-65b | 58.3 | 57.8 | 84.2 | 48.8 | 42.3 |
| pinkmanlove/llama-65b-hf | 58.3 | 57.8 | 84.2 | 48.8 | 42.3 |
| llama-65b | 58.3 | 57.8 | 84.2 | 48.8 | 42.3 |
| MetaIX/GPT4-X-Alpaca-30b | 57.9 | 56.7 | 81.4 | 43.6 | 49.7 |
| Arala/VicUnlocked-alpaca-30b | 57.6 | 55 | 88.8 | 44 | 50.4 |
| digitous/Alpaca30b | 57.4 | 57.1 | 82.6 | 46.1 | 43.8 |
| Arala/GPT4-x-AlpacaBento2-30b | 57.2 | 56.1 | 79.8 | 44 | 49.1 |
| TheBloke/dromedary-65b-lora-hf | 57 | 57.8 | 88.8 | 50.8 | 38.8 |
| TheBloke/Wizard-Vicuna-13B-uncensored-hf | 57 | 53.6 | 79.6 | 42.7 | 52 |
| elinas/llama-30b-hf-transformers-4.29 | 56.9 | 57.1 | 82.6 | 45.7 | 42.3 |
| austroba/llama30b-SuperHOT | 56.9 | 57.1 | 82.6 | 45.7 | 42.3 |
| llama-30b | 56.9 | 57.1 | 82.6 | 45.7 | 42.3 |
| cyl/awesome-llama | 56.8 | 54.4 | 79.7 | 41.8 | 51.3 |
| NousResearch/Nous-Hermes-13b | 56.4 | 52.5 | 88 | 41.8 | 51.1 |
| elinas/chronos-33b | 56.4 | 54.9 | 81.3 | 43.8 | 45.8 |
| HuggingFaceM4/starchat-beta | 56.2 | 52 | 88.6 | 44.7 | 47.3 |
| openaccess.ai-collective/wizard-mega-13b | 55.7 | 52.5 | 78.6 | 41 | 50.6 |



LLMS AS PART OF A SYSTEM

FOUNDATION MODELS

[2012-2022]

- **Apply DL to task X**
- End-to-End (Given inputs, predict outputs)
- Specific to one task
- Adaptation via fine-tuning only
- Examples
 - BERT
 - CNNs
 - ...

[2022- now]

- Combine foundation/base models to X,Y,Z to do task Q,R,S
- Can do many tasks
- Adaption via prompting (no parameter updating)
- Powerful building blocks
- Examples
 - LLMs
 - CLIP
 - ...

LLMS AS PART OF A SYSTEM

- ChatGPT and Chat UIs are fun, but basic
- The real power of LLMs comes as part of a system
- What are the building blocks?

Building Blocks

Memory

- Long term
- Short term

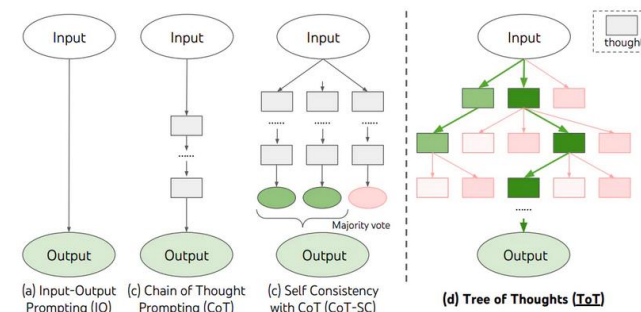
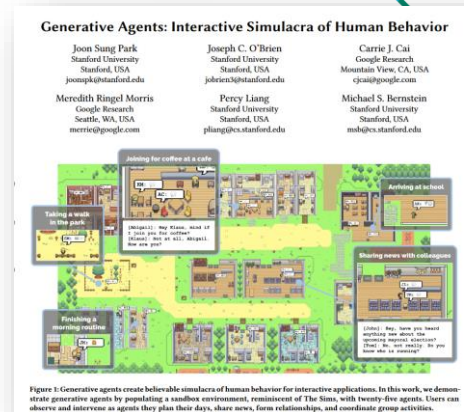
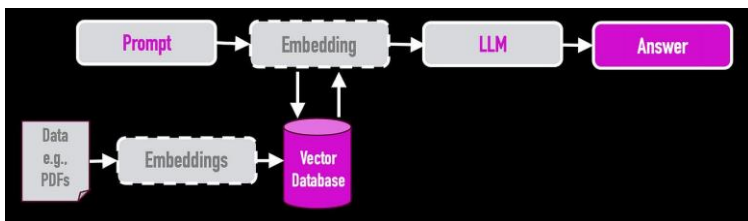
Tools

- Code interpreters
- Shells
- File system
- API

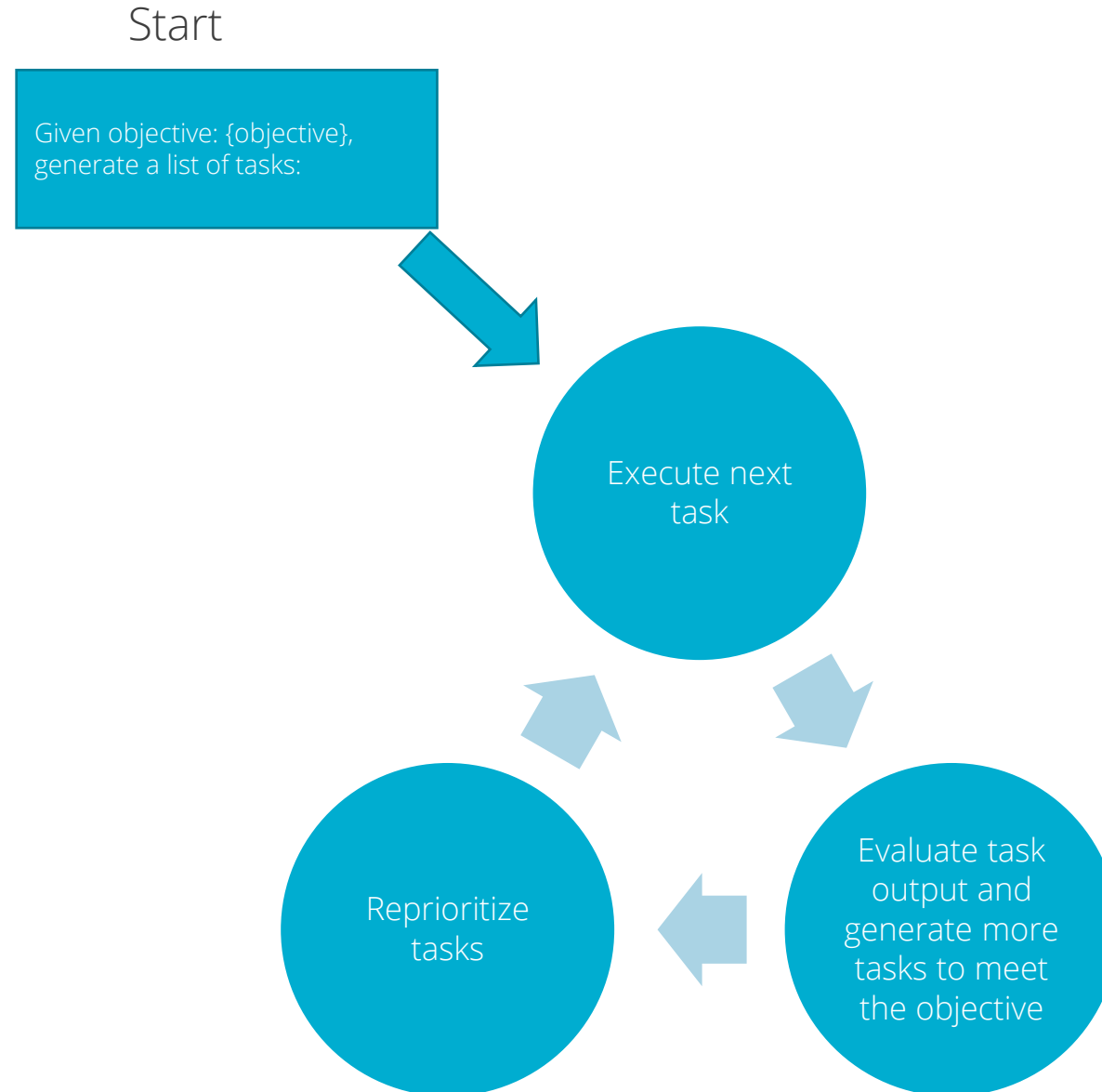
Database

LLM

- Tree of thought
- Reflection



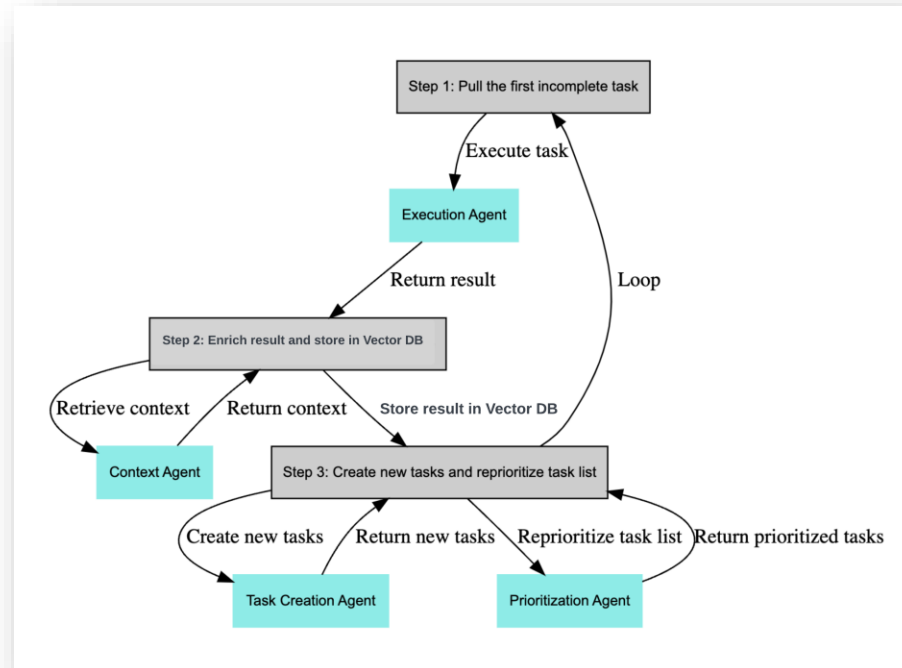
AGENTS



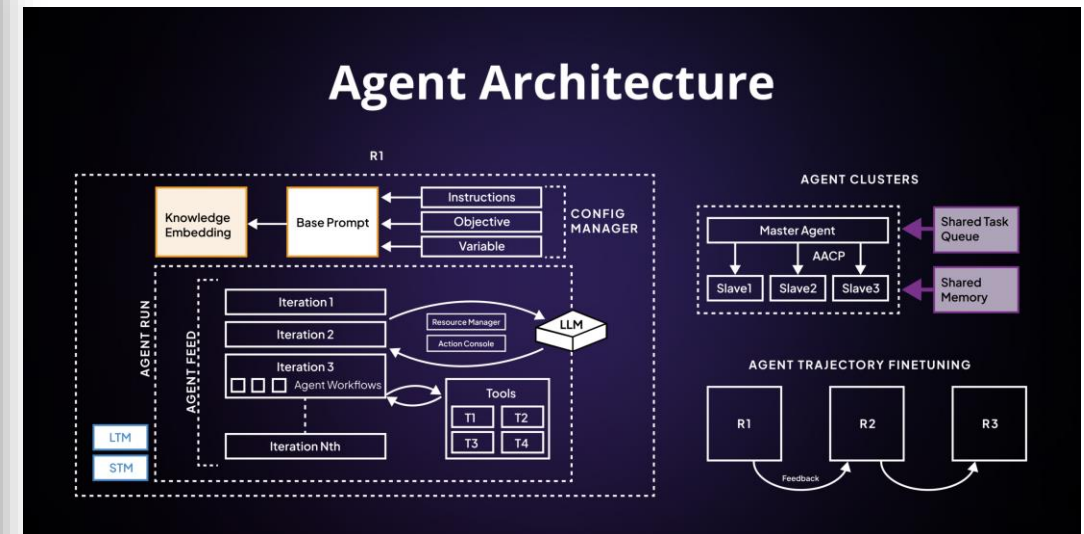
AGENTS

A few open source projects

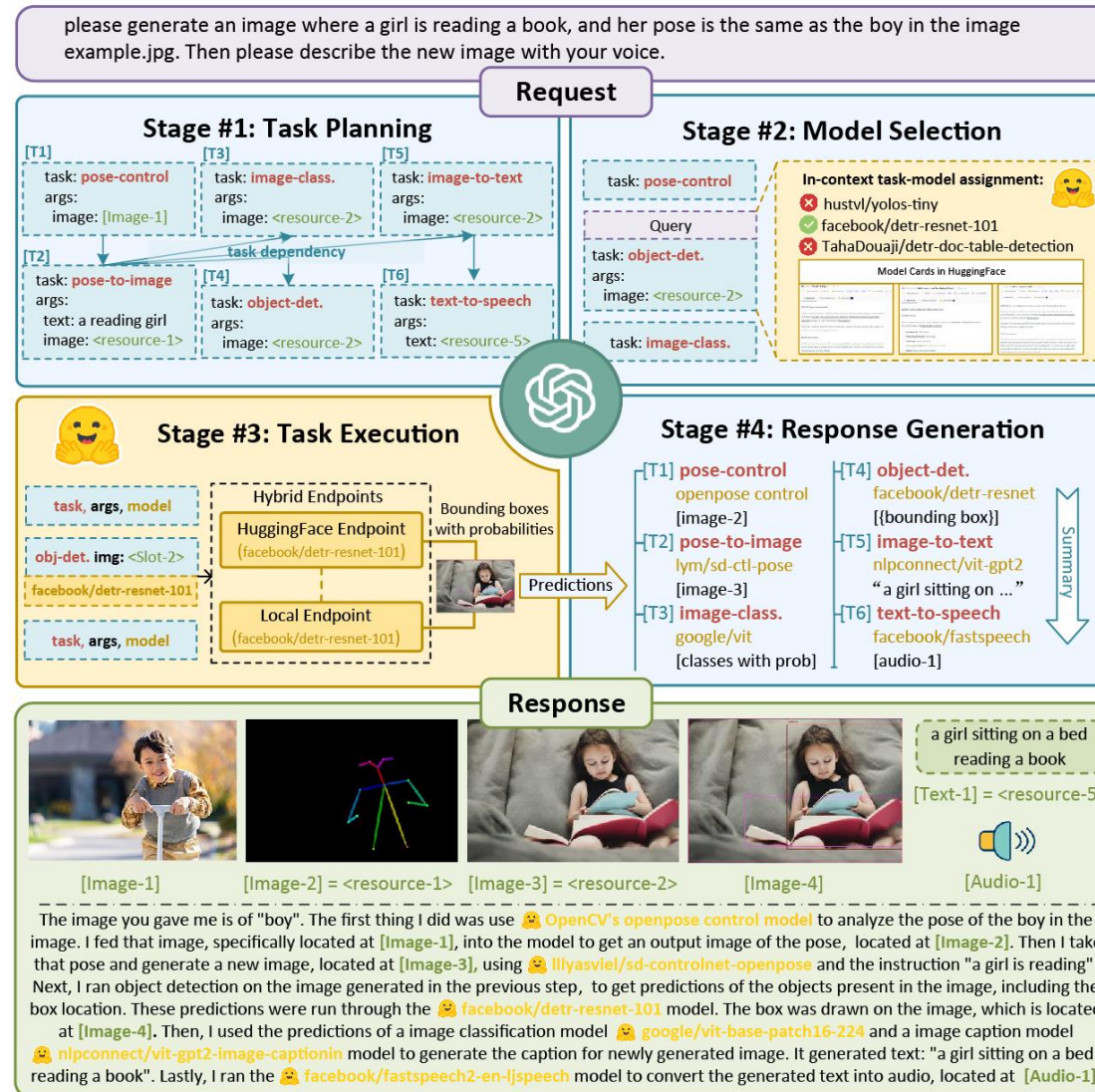
1. BabyAGI
2. AutoGPT
3. AgentGPT
4. SuperAGI
5. GPT-engineer
6. ...



Yoheinakajima BabyAGI



<https://github.com/TransformerOptimus/SuperAGI>



Write some code to make a
multiplayer game in python

<https://agentgpt.reworkd.ai/>

<https://github.com/reworkd/AgentGPT>

The screenshot displays the AgentGPT-3.5 interface with the following components:

- Task List (Left):** A series of tasks added to the project:
 - ★ Task Added: Define the gameplay mechanics and rules of the platformer game.
 - ★ Task Added: Design the levels and create the game assets (such as sprites and backgrounds).
 - ★ Task Added: Write the code for player movement, collision detection, and other gameplay features.
 - ★ Task Added: Test the game for bugs and make any necessary adjustments to ensure smooth gameplay.A "Starting task" is also shown: "Define the gameplay mechanics and rules of the platformer game."
- Current tasks (Right):** A sidebar showing the current task: "Define the gameplay mechanics and rules of the platformer game." It also lists other tasks: "Design the levels and create the game assets (such as sprites and backgrounds)", "Write the code for player movement, collision detection, and other gameplay features", and "Test the game for bugs and make any necessary adjustments to ensure smooth gameplay."
- Execution Status:** A green checkmark indicates "Executing: Define the gameplay mechanics and rules of the platformer game."
- Gameplay Mechanics and Rules:** A section titled "Overview" describing a platformer game and a list of aspects to consider:
 1. Character Movement
 2. Collision Detection
 3. Platform Generation
 4. Obstacle Generation
 5. Scoring System
 6. Level ProgressionIt then states: "Let's define each of these aspects in detail."
- Character Movement:** A sub-section titled "1. Character Movement" describing the character's movement capabilities and providing the following Python code:

```
class Character:
    def __init__(self):
        self.x = 0 # character's x position
        self.y = 0 # character's y position
        self.velocity_x = 0 # horizontal velocity
        self.velocity_y = 0 # vertical velocity

    def move(self, delta_time):
        self.x += self.velocity_x * delta_time
```
- Chat Interface:** A "Chat with your agent..." input field at the bottom.
- Controls:** A "Custom task" button and an "Add" button at the bottom right, and a "Copy Code" button next to the code block.

AgentGPT templates available now



| | | | |
|---|--|--|---|
| ResearchGPT
Academics and Professional
Generate a thorough report on a specific subject | BrandGPT
Academics and Professional
Evaluate a brand's performance, market position, and future prospects | TravelGPT
Other
Plan a detailed journey to a selected destination | PlatformerGPT
Creative and Social
Code a platformer game featuring a popular character or theme |
| IndustryGPT
Academics and Professional
Present a comprehensive review of an industry, covering key trends, players, and future predictions | ScraperGPT
Other
Extract and summarize data from a selected website | PostGPT
Creative and Social
Create engaging captions and hashtags for your social media posts | EmailGPT
Academics and Professional
Compose a concise and detailed email |
| ResumeGPT
Academics and Professional
Design a professional resume based on your career history and skills | NovelGPT
Creative and Social
Begin writing a novel in a selected genre | DietGPT
Health and Fitness
Create a customized diet plan based on dietary preferences and goals | FitnessGPT
Health and Fitness
Design a workout regimen tailored to your fitness goals |
| MarketingGPT
Academics and Professional
Design a comprehensive marketing strategy for your business | BudgetGPT
Academics and Professional
Prepare a personal or family budget | StudyGPT
Academics and Professional
Design a study schedule to achieve your academic objectives | NewsGPT
Other
Author a detailed news article on a selected topic |
| EventPlannerGPT
Other
Organize a detailed schedule for your forthcoming event | BlogGPT
Creative and Social
Write a blog post on a selected topic | AstroGPT
Science and Technology
Discuss astronomical phenomena, discoveries, and related technology | ArtReviewGPT
Creative and Social
Critique a piece of art, discussing its style, context, and influence |

SNL can make their own

| | | | |
|--|--|--|---|
| EngineerGPT
Assist an engineer in designing a 3D object | ScienceGPT
Analyze this data and write a report | SierraGPT
Design a sierra input deck based on these requirements and 3d models | Sierra2GPT
Interpret the output of sierra to see if it meets these requirements |
| ContractGPT
Design a WPA with another agent to get job X done. | SummaryGPT
Given these historical documents and these new concerns, write a report | | |

CONCLUSION

- Chatbots are cool, but aren't why LLMs are so powerful
- LLMs as a system are powerful
 - LLM
 - Controller
 - Synthesizer
 - Tools provide the system with more capabilities
 - Good systems are Self-correcting
- The exact setup of how to piece together different building blocks is an open question
- All the tools exist and open source right now!
 - Open source LLMs
 - Open source tool kits
- How to get started?
 - LangChain
 - LlamaIndex
 - AgentGPT
- Agents will only get better with time and research

